Signal Processing 89 (2009) 1038–1049

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Verified speaker localization utilizing voicing level in split-bands

Afsaneh Asaei [a,b,c,*], Mohammad Javad Taghizadeh [c], Marjan Bahrololum [c], Mohammed Ghanbari [d]

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland
[c] Iran Telecommunication Research Center, Tehran, Iran
[d] Department of Computing and Electronic Systems, University of Essex, Colchester, UK

ARTICLE INFO

Article history:
Received 26 December 2007
Received in revised form
31 October 2008
Accepted 4 December 2008
Available online 24 December 2008

Keywords:
Microphone array
Speaker verification
Speaker localization
Reverberation
Beamforming
Speech recognition

ABSTRACT

This paper proposes a joint verification-localization structure based on split-band analysis of speech signal and the mixed voicing level. To address the problems in reverberant acoustic environments, a new fundamental frequency estimation algorithm is proposed based on high resolution spectral estimation. In the reconstruction of the distorted speech this information is utilized to reduce the side effect of acoustic noise on the voicing parts. A speaker verification system examines the features of the reconstructed speech in order to authorize the speaker before localization. This procedure prevents localization and beamforming for non-speech and specially the unwanted speakers in multi-speaker scenarios. The verification is implemented with the Gaussian Mixture Model and a new filtering scheme is proposed based on the voicing likelihood of each frequency band measured in the previous steps for efficient localization of the authorized speaker. The performance of the proposed VSL (verified speaker localization) front-end is evaluated in various reverberant and noisy environments. The VSL is utilized in the development of distant-talking automatic speech recognition by microphone array where the system can lock on a specific source and hence the recognition quality improves noticeably.

1. Introduction

For a hands-free speech interface, it is very important to capture distant talking speech with high quality. An ideal solution for this purpose is sound acquisition by microphone array. A microphone array can acquire the desired speech signals selectively by steering the beam pattern directivity of the array towards the desired speaker. This process is called beamforming and due to the directivity pattern steering, it can spatially filter out noises from other directions regardless of the noise nature. The main obstacles to achieve reasonable performance in array based systems are the reverberation and the presence of ambient noise of acoustic environment. These parameters affect the accuracy of speaker localization and beamforming in capturing the desired spatial signal and suppressing the others. To tackle this problem, various methods have been proposed recently, but they all seem to give erroneous estimations in speaker direction finding under the presence of high noise and reverberation. These conventional algorithms in multi-speaker environments not only have difficulty in localizing the multiple sound sources accurately, but they also fail to localize the target talker among the known multiple speaker positions. These localization techniques can be loosely classified into three general categories:

* Corresponding author at: IDIAP Research Institute, Martigny, Switzerland. Tel.: +41 27 721 77 73.
E-mail addresses: afsaneh.asaei@idiap.ch (A. Asaei), taghizadehmj@itrc.ac.ir (M.J. Taghizadeh), bahrololum@itrc.ac.ir (M. Bahrololum), ghan@essex.ac.uk (M. Ghanbari).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.

(i) those adopting high resolution spectral concepts, (ii) techniques based upon maximizing the steered response power (SRP) of a beamformer and (iii) approaches employing time-difference of arrival (TDOA) information.

The first class of these techniques, characterizes any localization scheme that is dependent upon applications of the spatio-spectral correlation matrix [1]. Interestingly, all of these methods are all designed for narrowband signals and are very sensitive to source and microphone modeling [2] implying complexities within the speaker localization process [3,4]. The second class of the aforementioned strategies is based on maximizing the output power of a steered beamformer or SRP. In this case, a beamformer is used to scan over a predefined spatial region by adjusting its steering delays [5]. A filtering process can also be employed to increase accuracy whereby filters are designed in such a way to boost the power of the desired signal even if they may increase distortion. This is the main distinction between the popular beamforming techniques in speech acquisition systems and that of localization [6,7]. This category has the most robustness in source localization in practical situations and is preferable in enabling reliable localization of speech signals with short frames [8]. The third category is realized in two phases. Firstly, it detects a set of TDOA of the wave-front between different microphone pairs mostly based on the generalized cross-correlation (GCC) function maximization [9]. In computing the cross-correlation function, to increase accuracy, some weighting schemes are also employed. The most important weightings are ML (Maximum Likelihood) and PHAT (phase transform) [10,11]. Second, geometrical constraints are used to infer the source position. Due to its low computational cost, this technique has attracted many interests. However, pair-wise techniques suffer considerably from multipath propagation [8]. Since the primary goal of microphone array based systems is practicality in the real environment, we have considered this subject for real applications. In the scenario which is the subject of this investigation, we have focused on SRP based localization.

All the above mentioned attempts were aimed to improve the localization accuracy in the presence of acoustic noise and reverberation and could not achieve satisfactory results in the presence of spurious speech sources such as the voice of unwanted speakers. In this scenario, speaker verification is needed to authorize the speech. This stage of speaker verification by microphone array is addressed in [12], where a microphone array is utilized to capture the speech and provide input for automatic speech identification. A 2-D matched filter microphone array is proposed to improve the identification scores in a reverberant environment. In this algorithm, the identification is addressed after the array-based analysis of the received signal. Investigations by Gianakopoulos et al. [13] are concentrated on the implementation of the front-end signal pre-processing tasks such as filtering, acquisition and beamforming to improve speaker recognition. This procedure suffers from over computation scenarios. In [14] an adaptive near-field beamformer is implemented for hands-free speaker recognition. In [15] speech enhancement techniques are utilized to reduce the acoustic degradation of source signal and improve speaker verification in the noisy environments. In [16] a speaker identification algorithm based on the angle of arrival of the speech is proposed. Since the convergence rate is large, the new algorithm has practical limitations and participants are required to remain seated during the experiment. Hence, limited number of investigators has studied speaker recognition and although the effectiveness of beamforming is proven in robust hands-free speaker recognition [17], but verification always comes after the localization, beamforming and other computational array processing algorithms.

In this paper, the idea of verification prior to localization is proposed. It has been observed through extensive testing that the quality of the voiced parts is very important for verification. Therefore, we have enhanced these parts and used them for verification. For the verified speech, localization is performed and the enhanced signal is acquired through sub-array beamforming. The verification result is tested again after beamforming to ensure a high accuracy. We name this front-end block as verified speaker localization (VSL). The multi-channel speech enhancement based on localization and beamforming is only run for the desired voices and the whole system becomes robust to unwanted noises as well as other spontaneous sources of energy. The over computation of beamforming and post processing for unwanted speech signals is also prevented which reduces the computational complexity of the front-end task in multi-speaker scenarios considerably.

Organization of the paper is as follows: The general architecture of the proposed VSL front-end is explained in Section 2. It includes a brief overview of VSL components, details of the split-band reconstruction, speaker verification and localization. Scenario of testing and the results achieved are described in Section 3. A VSL based far-field automatic speech recognition (ASR) is also introduced in this section and the effect of the VSL front-end on the performance of this system is evaluated. Finally, concluding remarks are given in Section 4.

2. General overview of the proposed VSL front-end

The main elements of the proposed front-end signal pre-processing block are: acquisition, reconstruction of the voiced parts, verification, localization and beamforming. The order in which they interact with each other is shown in Fig. 1.

The acquired speech is first analyzed in split-bands to measure the voicing level. For this purpose in the reverberant acoustic environments, a new fundamental frequency estimation algorithm is proposed based on the subspace approach in high resolution spectral estimation. A reconstruction stage for the degraded voiced bands is also proposed prior to the verification. The verification is implemented using Gaussian Mixture Model and a new

Fig. 1. A general architecture of the proposed VSL front-end.

likelihood of each frequency band measured in the previous steps to effectively localize the authorized speaker.

In the traditional methods, as discussed in the introductory part, whenever a source of energy is detected by the localization algorithm, the beamforming will then be applied to acquire the enhanced signal. These two processes are computational intensive in the far-field interfaces. In the proposed VSL front-end, a new localization algorithm improves the speaker localization accuracy as well as the robustness against the reverberation and noise, while the verification which is performed prior to localization prevents the over computation of localization and beamforming for unwanted sources (specially transient or unauthorized speakers). Therefore, the whole system will have the capability to update the location information of any specific individual. On the other hand, since the localization is based on short speech frames, it is also capable of tracking a moving speaker. These two capabilities indicate that the system can lock on a speaker, while ignoring other speech sources. Since localization and beamforming are highly computational demanding [11] and achieving an enhanced speech for far-field applications needs heavy processing, this lock on characteristic improves the front-end task both in terms of computation and robustness in far-field applications such as teleconferencing, voice control and speech recognition where the presence of unwanted speech signals is highly probable.

In the proposed VSL front-end, the received signal is first segmented based on detection of the non-speech activity for more than 2 s. Each segment is analyzed for voicing level measurement at speech sub-bands corresponding to the fundamental frequency harmonics. The voiced parts are then reconstructed at split bands regarding the harmonic bands of the speech spectrum and the signal is analyzed for authentication within a verification algorithm. For the verified speech, misdetection of source localization due to reverberation and acoustic noise is reduced through the voicing level measurement. The beamforming algorithm uses this information to steer the beam pattern towards the direction of the speaker to acquire the source signal while suppressing the noise from other directions. Details of each component are discussed in the following sections.

2.1. Microphone array signal model

In this paper, we assume the sound wave propagation follows a linear wave equation [18]. Hence, the acoustic path between the sound sources and microphones can be modeled as a linear system [19]. This assumption is plausible in small-room microphone array environments and is usually employed in the array-processing techni-

the mth microphone at location d_m can be expressed as

$$x_m(t) = s(t) * h_s(d_m, d_s, t) + v_m(t) \quad (1)$$

where $h_s(d_m,d_s,t)$ is the room impulse response from the speech source $s(t)$ at location d_s to microphone m. The operator $*$ is convolution. v_m is a white Gaussian and is assumed to be uncorrelated to $s(t)$.

The impulse response h, characterizes all the acoustic paths from the source to location d_m, including the direct path. In general, h_s varies with environmental changes, such as temperature, humidity, furniture and people inside the room. It is reasonable to assume these factors to remain fixed in the period of each experiment. Separating the direct path component from the rest of the acoustic paths, the following expression can be defined for $h_s(d_m,d_s,t)$:

$$h_s(d_m, d_s, t) = \frac{a}{r_m}\delta(t - \tau_m) + u(d_m, d_s, t) \quad (2)$$

where r_m is the distance between the source and the mth microphone, τ_m is the propagation delay equal to the ratio of r_m to the speed of sound. The constant a depends on the medium and the system of units used. $u(d_m,d_s,t)$ characterizes all the acoustic paths except the direct path. Substituting this equation into (1), the signal model at microphone m is given by

$$x_m(t) = \frac{a}{r_m}s(t - \tau_m) + s(t) * u(d_m, d_s, t) + v_m(t) \quad (3)$$

The first term is the direct path component which is important for localization, the second term is the model of reverberation and the third term is the uncorrelated noise.

2.2. Split-band reconstruction

A typical simulated room impulse response is illustrated in Fig. 2. The largest peak corresponds to the direct path and the other peaks are due to the surrounding walls reverberation. Assuming the total system of microphone array and room as a linear system [21], the received signal at each microphone is the convolution of this impulse response with the original source signal. This effect impairs the received signal quality at the microphone array and reduces the periodicity of the voiced segments. Hence we have considered this side-effect and have enhanced these harmonic parts through reconstruction.

The first step is the estimation of the fundamental frequency. However, due to the distortion of periodicity and harmonicity, conventional fundamental frequency extraction algorithms such as autocorrelation function (ACF), average magnitude difference function (AMDF), Cepstrum, simple inverse filtering tracking (SIFT) and harmonic product spectrum (HPS) give erroneous results. Since the estimation accuracy of the fundamental frequency in the presence of noise and reverberation is very

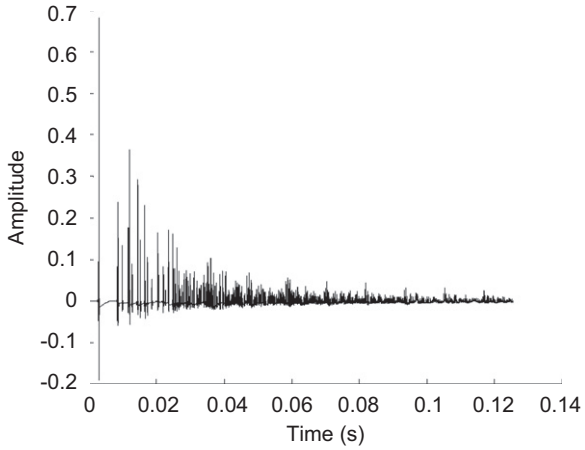


Fig. 2. Room impulse response.

have extracted the fundamental frequency on the subspace to benefit from the high resolution spectral estimation property of this technique.

The subspace based spectral estimation is an accurate method for detecting the discrete frequencies of a signal and hence we used the multiple signal classification (MUSIC) [22,23] in our algorithm. The MUSIC algorithm detects complex sinusoids by performing eigendecomposition on the data vector covariance matrix of the received signal. Andrews et al. [24] have already proposed the pitch determination algorithm based on MUSIC. Here we have modified their approach for the reverberant signals. To find the fundamental frequency, the autocorrelation matrix of the speech signal is computed from its power spectrum via FT. Since the fundamental frequency of speech sources is less than 800 Hz [25], we have applied the MUSIC algorithm only to the lower frequency components of the speech spectrum. With an 800-point DFT of 20 ms of the speech signal at the sampling frequency of 16 kHz, the frequency components of a MUSIC spectrum will be at 20,40,…,800 Hz. The total number of these components is 40 and the eigenvalues are computed from the received signal autocorrelation matrix. The number of harmonics contained in the spectrum is an important parameter of the MUSIC algorithm. If it is set too large, the spectrum will be easily affected by the noise and if it is too small, the spectral estimation becomes inaccurate and the error will be increased. For our experiments, the set of dominant eigenvalues $\{\lambda_k\}$ which span over the signal subspace are chosen so as to satisfy $\lambda_1 \geqslant \lambda_k \geqslant \lambda_1/8$, where λ_1 is the eigenvalue of the first fundamental component. The FFT is applied to the logarithm of the MUSIC power spectrum and the peak location of the signal determines the estimated fundamental frequency. To reduce the computational cost, we have estimated the fundamental frequency at the precision of 20 Hz. This was done by searching the pseudospectrum of the signal with 1 Hz precision at the vicinity of 80 Hz around the pre-estimated fundamental frequency. The corresponding frequency of the local maxima is detected as the fundamental

Since the room can be modeled as a linear system, the frequency content of the received signal is similar to the original sound and it is only distorted in amplitude and phase. Therefore reverberation converts the global maximum of the spectrum to a local maximum with no frequency displacement.

Through a large number of experiments we have verified the robustness of the algorithm to different reverberant noisy environments. The algorithm was also verified for robustness to sudden closure, such as in a vowel-to-nasal transition, where waveform periodicity is reduced but the fundamental frequency did not change.

After estimation of the fundamental frequency, the algorithm is used to measure the voicing level in each frequency band. An accurate measure of voicing level was applied to multi-band excitation (MBE) coders [26]. The voicing decision was made by calculating the normalized error E_l between the original and the modeled speech spectrum in each frequency band of the fundamental frequency harmonics:

$$E_l = \frac{\sum_{\omega=a_l}^{b_l} |X(\omega) - \hat{X}(\omega, \omega_0)|^2}{\sum_{\omega=a_l}^{b_l} |X(\omega)|^2} \tag{4}$$

where $X(\omega)$ is the speech spectrum of the received signal at the reference microphone channel (#5), ω_0 is the fundamental frequency, a_l and b_l are the first and last harmonics in the lth band, and $\hat{X}(\omega, \omega_0)$ is the estimated speech spectrum calculated in each frequency band as the spectral shape of a Hanning window with a constant amplitude.

To determine the voicing decision, the normalized error, E_l, of the lth frequency band is compared with an adaptive threshold [27]. If the normalized error is less than a threshold, the corresponding frequency band belongs to the target voice and it is reconstructed in the split-bands based on the fundamental frequency harmonics.

Since higher harmonics are more susceptible to reverberation and acoustic noise [28] decision on voicing for the frame was carried out on the majority of the lower half of the speech frequency band. For those intervals when all of the speakers are talking simultaneously, the speech frames loose their periodicity and these frames are not involved in the other phases of the VSL processing.

The speech signal due to acoustic noise is distorted. The distortion can be reduced in voiced parts by precise extraction of the fundamental frequency and then using it to reconstruct the speech spectrum. The split-band mixed voicing decision calculated for each frequency band is utilized to synthesize the voiced speech spectrum. Each harmonic band has a shape similar to the spectral shape of the window used prior to the Fourier transform, whereas the non-voiced bands are random in nature. Therefore, a voiced harmonic band can be finely synthesized as a multiplication of the frequency response of a suitable window centered at the harmonic of fundamental frequency corresponding to that band with constant amplitude measured with respect to the original signal [29].

Reconstruction of the harmonic bands is given by

voiced band of the speech spectrum.

$$\hat{X}(\omega,\omega_0)=A_{k,\omega_0}W(\omega)\quad 1\leqslant k\leqslant K,\ \ \lceil a_k\rceil\leqslant\omega\leqslant\lceil b_k\rceil \tag{5}$$

where $a_k=(k-0.5)\omega_0$, $b_k=(k+0.5)\omega_0$, $\lceil\cdot\rceil$ stands for the nearest integer greater than or equal to, K is the number of harmonics in the 8 kHz speech frequency bandwidth, $W(\omega)$ is the frequency response of the Hanning window centered at the kth harmonic of the fundamental frequency and A_{k,ω_0} is the kth harmonic amplitude defined as:

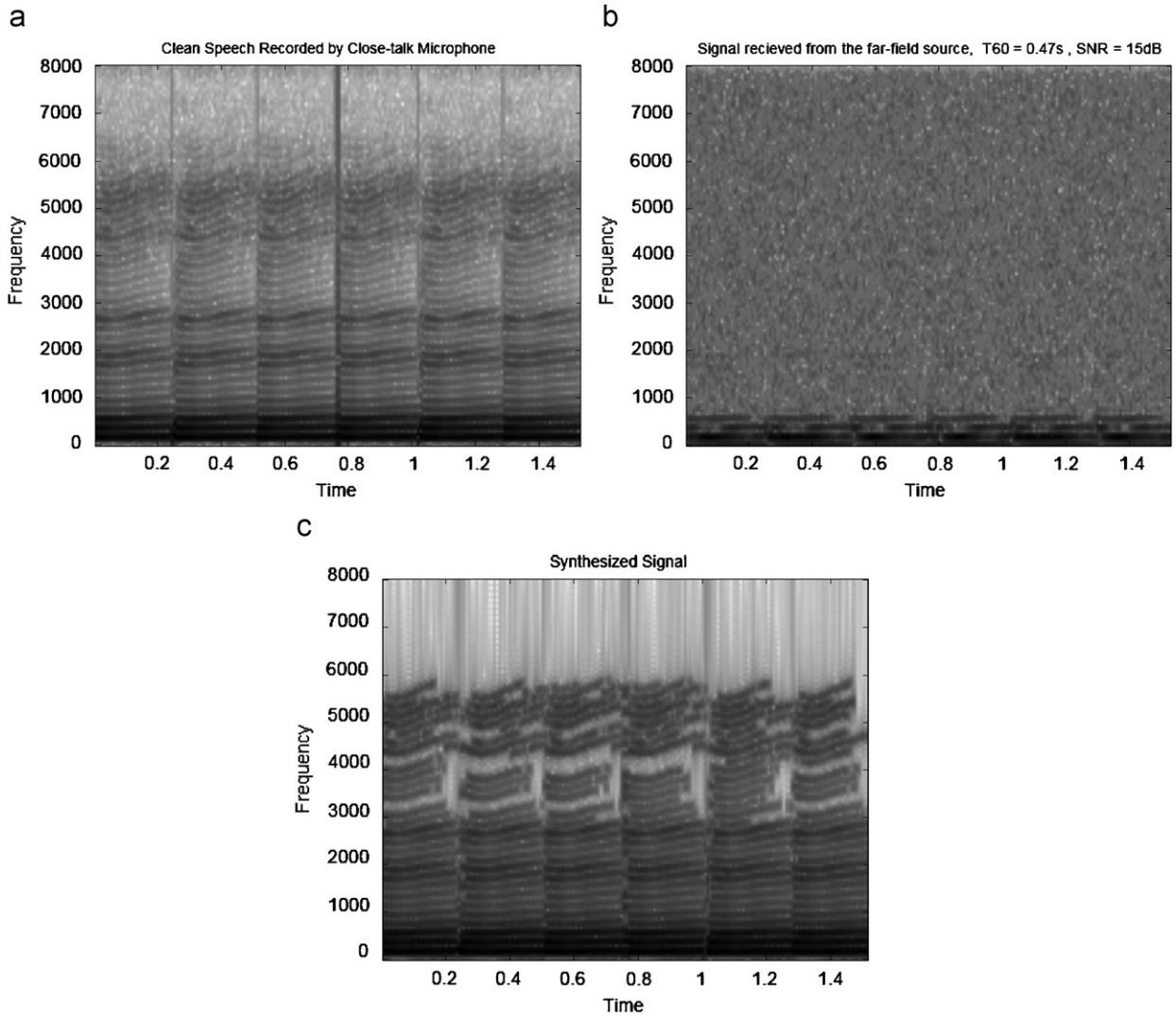
$$A_{k,\omega_0}=\frac{\sum_{\omega=\lceil a_k\rceil}^{\lceil b_k\rceil}X(\omega)W(\omega)}{\sum_{\omega=\lceil a_k\rceil}^{\lceil b_k\rceil}|W(\omega)|^2} \tag{6}$$

For concatenation of the reconstructed successive frames, we use linear interpolation to remove frequency mismatches [30]. Fig. 3 displays a clean speech, noisy signal and the synthesized speech from its noisy origin by spectrogram. This figure shows how reconstruction procedure reduces the acoustical noise and retrieves the harmonicity of voicing speech.

2.3. Speaker verification

Mixture models belong to a family of density model that comprises of a number of component functions, usually Gaussian. The distribution of feature vectors was extracted from a speaker's speech modeled by a Gaussian mixture density. This is a method that has been proven to be one of the most successful approaches for text-independent speaker verification. Therefore we have implemented speaker modeling based on the Gaussian Mixture Models (GMM). In this algorithm Gaussian mixtures are used to model arbitrary densities of the speech signal [31–33].

A block diagram of the implemented speaker verification system is shown in Fig. 4. There are two steps in the



DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.